Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study

Year: 2023 | Citations: 130 | Authors: Kostis Giannakopoulos, Argyro Kavadella, Anas Aaqel Salim, Vassilis Stamatopoulos, E.

Abstract

Background The increasing application of generative artificial intelligence large language models (LLMs) in various fields, including dentistry, raises questions about their accuracy. Objective This study aims to comparatively evaluate the answers provided by 4 LLMs, namely Bard (Google LLC), ChatGPT-3.5 and ChatGPT-4 (OpenAI), and Bing Chat (Microsoft Corp), to clinically relevant questions from the field of dentistry. Methods The LLMs were queried with 20 open-type, clinical dentistry–related questions from different disciplines, developed by the respective faculty of the School of Dentistry, European University Cyprus. The LLMs' answers were graded 0 (minimum) to 10 (maximum) points against strong, traditionally collected scientific evidence, such as guidelines and consensus statements, using a rubric, as if they were examination questions posed to students, by 2 experienced faculty members. The scores were statistically compared to identify the best-performing model using the Friedman and Wilcoxon tests. Moreover, the evaluators were asked to provide a qualitative evaluation of the comprehensiveness, scientific accuracy, clarity, and relevance of the LLMs' answers. Results Overall, no statistically significant difference was detected between the scores given by the 2 evaluators; therefore, an average score was computed for every LLM. Although ChatGPT-4 statistically outperformed ChatGPT-3.5 (P=.008), Bing Chat (P=.049), and Bard (P=.045), all models occasionally exhibited inaccuracies, generality, outdated content, and a lack of source references. The evaluators noted instances where the LLMs delivered irrelevant information, vague answers, or information that was not fully accurate. Conclusions This study demonstrates that although LLMs hold promising potential as an aid in the implementation of evidence-based dentistry, their current limitations can lead to potentially harmful health care decisions if not used judiciously. Therefore, these tools should not replace the dentist's critical thinking and in-depth understanding of the subject matter. Further research, clinical validation, and model improvements are necessary for these tools to be fully integrated into dental practice. Dental practitioners must be aware of the limitations of LLMs, as their imprudent use could potentially impact patient care. Regulatory measures should be established to oversee the use of these evolving technologies.